

# Integer Set Compression and Statistical Modeling

N. Jesper Larsson

IT University of Copenhagen, Denmark,  
jesl@itu.dk

**Abstract.** Compression of integer sets and sequences has been extensively studied for settings where elements follow a uniform probability distribution. In addition, methods exist that exploit clustering of elements in order to achieve higher compression performance. In this work, we address the case where enumeration of elements may be arbitrary or random, but where statistics is kept in order to estimate probabilities of elements. We present a recursive subset-size encoding method that is able to benefit from statistics, explore the effects of permuting the enumeration order based on element probabilities, and discuss general properties and possibilities for this class of compression problem.

## 1 Introduction

Data compression in its most basic form is commonly expressed in terms of representing a string of characters drawn from a fixed alphabet. A situation with somewhat different characteristics is when the data to represent is a *set*, i.e., a sequence of non-repeating elements whose order is insignificant. Although it is possible to transform the one scenario to the other and vice versa, probability distributions and applicable modeling schemes differ, and there is benefit in treating the problems separately.

This work focuses on compression of sets whose elements are drawn from a fixed range of integers, which we refer to as the *universe*. Another interpretation, common in related work [12,19,20] is to view the items to be compressed as the differences between consecutive elements in sorted order, i.e., a sequence of integers. The same encoding methods can be described in terms of set or integer sequence compression [12]. In this work, we prefer the set interpretation, since we are interested in using statistics for individual elements of the universe (as opposed to the gaps between them).

Compression of sets has a number of uses as a component of other compression or data structure problems. One of the more prominent ones is storage of *inverted indexes* [22,21]. Others appear in a wide variety of applications such as succinct data structures [14], data mining [18], and web graph representation [1]. Foundations for coding of sets go many decades back [4,6,3] and developments stretch into recent work. Of particular interest as related to this work are interpolative coding and related methods [13,20,19] and methods that use binary tries for compressing sets and multi-sets [15,7]. There are, however, different classes of modeling assumptions, and works are not generally applicable to the same settings and applications. In particular, little work has been published that attempts to make use of statistics over elements, which is among our main focal points.

This work is outlined as follows. Section 2 relates previous methods of particular importance to our work. We note that a slight optimization of cap coding is possible for the fixed-universe set compression. Section 3 presents our method of *recursive subset-size*, relates it to other methods, and discusses statistical set compression issues in general. Section 4 concludes and points to future research. Parts of this work have been previously presented in poster form [10].

*Formal Problem and Notation* In general, data compression can be expressed as encoding a message into a compact format by which it can subsequently be reconstructed by a decoder, using a set of code-specific premises shared by encoder and decoder. We view the encoding process as a sequence of *emit* operations, which each specify an event corresponding to a property of the message. An emit contributes a number of bits to the encoded output. Ideally, emitting an event that has probability  $p$  should take  $-\log_2 p$  bits [17]. Given that probability ranges of the possible events to be emitted can be inferred in the same way by encoder and decoder, we can use arithmetic coding [16] to produce a number of bits arbitrarily close to the ideal, even when the desired number of bits is a fractional number or less than one. Hence, we generally assume that ability to estimate probabilities is enough to uniquely define both encoding and decoding.

The special case of encoding an integer  $x$  such that  $L \leq x \leq H$  for integers  $L$  and  $H$ , we denote as emitting  $x[L, H]$ . The bits thus produced depend on the encoding used, and may also depend on probability estimates for the numbers  $L, \dots, H$  shared by encoder and decoder. When  $L = H$ , zero bits are produced.

We study the problem of encoding a set  $S$  consisting of  $|S|$  integers drawn from universe  $U = 0, 1, \dots, |U| - 1$ . An equivalent interpretation is to view elements as bitstrings whose lengths are limited by  $\lceil \log_2 |U| \rceil$ . We use these interpretations interchangeably. We assume that knowledge of  $|U|$ , which completely defines  $U$ , is shared by encoder and decoder. Although we do not generally consider  $|S|$  to be known in advance, we do not devote much effort to the encoding of  $|S|$ . Most of the methods we consider (the only exception being the yes/no code in section 3.1) depend on  $|S|$  being encoded separately, and its choice of code is independent of the main coding scheme. Section 3.2 does, however, address encoding of  $|S|$ .

*Note on Experiments* This work is not directed at any particular application area. In order to evaluate performance, we test on primarily three instances of natural data, with different characteristics consisting of small and moderate-sized sets, as well as on some extreme generated data. The first natural data instance, *txt*, tests performance on a very small universe. It takes elements as bits of either individual characters ( $|U| = 8$ ) or three bytes grouped together ( $|U| = 24$ ). The other two sets are generated from a set of Unix documentation files. In one, *words*, the sets are files, and elements are randomly assigned numbers of the words contained in the set. In the other, *inverted*, each set corresponds to the numbers of the files (randomly assigned) in which the word appears. For *words*,  $|U| = 19515$  and average  $|S|$  is 634. In *inverted*,  $|U| = 337$  and average  $|S|$  is 11.

## 2 Gap and Range-Narrowing Codes

This section describes previous methods of particular relevance to our work. Gap coding is the classic methods for independent elements. Range-narrowing methods recursively encode elements, and perform particularly well for clustered elements.

### 2.1 Gap Codes

Set representation can be transformed to sequence representation by arranging the elements of  $S$  in increasing order, and representing a sequence of gaps between adjacent elements. This is a common technique, described comprehensively e.g. by Witten, Moffat, and Bell [21].

Assuming that  $|S|$  is encoded separately before the elements, and that all elements are equally likely, we have, for a specific  $S$  and any  $x \in U$ , a global probability  $p = \Pr(x \in S) = |S|/|U|$ . Hence, the probability of gap size  $k$  can be estimated by the geometric distribution [9] as  $(1-p)^{k-1}p$ . Computing probability ranges in accordance with this distribution, we can achieve minimal encoding length with arithmetic coding [16], or a Golomb code [6] that approaches the same property.

We note, however, that geometric distribution is an approximation, corresponding to draws *with replacement* from a set of size  $|U|$  with  $|S|$  success states. In actuality, since elements in a set are distinct, the draws are *without replacement*. Taking this into account yields a slightly better estimate. Let  $V \subseteq U$  be the part of  $U$  that remains after encoding  $|S| - n$  elements. Then the probability of the next gap size being  $k$  is  $\prod_{i=0}^{k-1} (1 - n/(|V| - i)) \cdot n/(|V| - k)$ . For small  $|U|/|S|$  this can yield a noticeable difference, as seen can be seen on the *txt* data results in table 1. A similar argument can be used for modifying Golomb [6] or Elias codes [4] to reflect that numbers are chosen from a limited, decreasing, range.

### 2.2 Range-Narrowing Codes

Interpolative coding [13,20] uses a *low-short* binary code [19] to encode first the highest-numbered element, and then the median element of  $S$ . It then progresses recursively in the subsets below and above the median, always encoding the median, as deeply as necessary to uniquely represent every element. The size of the set is represented separately.

In terms of compression ratio, the strength of interpolative coding is that if the elements of  $S$  are clustered (i.e. have numbers close together), recursive progression quickly narrows in on small subsets of  $U$ , requiring only a few bits for each binary code.

The closely related *tournament coding* [19] is formulated as compression of an integer sequence, corresponding to the differences between consecutive set elements in sorted order. It progresses recursively over the sequence, encoding in each step the maximum element in the range. The original version of tournament coding works for unlimited-size integers, and the global maximum is submitted using Elias' gamma code [4]. In our range-limited setting within a known  $|U|$ , the maximum is better encoded using the same high-short binary code as the rest of the elements. In our tests, the results are roughly similar to those of interpolative coding, over which Teuhola demonstrates an advantage for uniform distributions.

### 3 Recursive Subset-Size Code and Use of Statistics

Gap-oriented methods adapt only to the *global* density of elements, based on a single set-size parameter. Range-narrowing methods are able to exploit *local* density differences, by reducing the codeword length for elements. But neither of the methods presents a natural way of exploiting statistical data about the frequencies of individual elements. We now consider coding schemes that do, to varying degrees.

#### 3.1 Prelude: Yes/No Code and Exponential Statistics

Assume that we can predict the probability for the inclusion of every possible element being included in the set to encode, i.e., for every  $x \in U$  we have an estimate of  $\Pr(x \in S)$ . Then arithmetic coding lets us emit  $|U|$  *included* or *not included* events, one for each possible element, using the corresponding probability range, by which we obtain a total encoded length of the optimal  $-\sum_{x \in U} \Pr(x \in S)$ .

Although this is optimal if inclusion in the set is independent among the possible elements, it ignores any correlation between elements. For example, say that two elements  $x$  and  $y$  usually appear together, i.e.,  $\Pr(x \in S \wedge y \in S) > \Pr(x \in S) \times \Pr(y \in S)$ . We could address this by keeping statistics on element probability conditioned on inclusion or exclusion of the previous elements in the yes/no encoding order. (By basic laws of conditional probability [9], the order has no impact on the overall probability estimate of a specific set, and hence neither on the optimal encoding length.) However, this would require statistics whose storage space is exponential in  $|U|$ , and is hence only realistic for small universes.

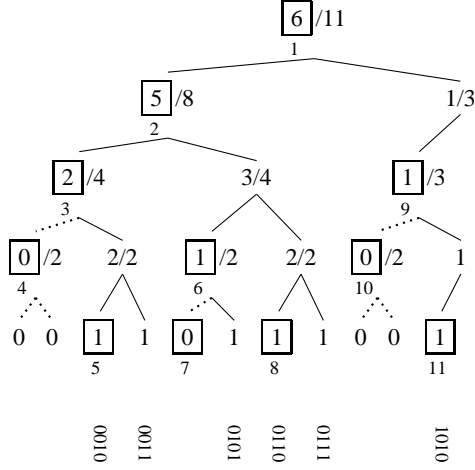
In principle, since the yes/no code is neutral as for how  $\Pr(x \in S)$  is estimated, it can be used for generating an optimal encoding length given any probability model. However, the neutrality also implies a lack of support for efficiently implementing any particular model. Furthermore, it always requires  $|U|$  emissions, which is not efficient for small sets drawn from a large universe. Table 1 includes encoding lengths for the yes/no code with globally calculated probability estimates as a baseline comparison for the other methods.

The yes/no code can be expected to produce the same encoding length as a Huffman code over the  $2^{|U|}$  possible sets, where the probability of a specific set  $\tilde{S}$  is  $\prod_{x \in \tilde{S}} \Pr(x \in S) \times \prod_{x \notin \tilde{S}} 1 - \Pr(x \in S)$ . Again, this is a construct exponential in  $|U|$ , and hence only viable for small universes.

#### 3.2 Recursive Subset-Size Code

We now present a code that recursively emits subset sizes over the left and right half of the element range, which we refer to as RSSS. This is somewhat similar to the range-narrowing codes, but allows the use of individual probability estimates, including a certain degree of context information. The number of counters for maintaining statistical information is bounded by  $|U|$ , a compromise with feasible space requirements even for large universes.

We begin by describing the basic method without use of statistics.



**Fig. 1.** Binary tree illustrating RSSS coding of  $S = \{0010, 0011, 0101, 0110, 0111, 1010\}$  from universe  $U = \{0000, \dots, 1010\}$ , i.e.,  $|U| = 11$ .

Consider the binary tree over  $U$  where  $|U| = 11$ , as shown in figure 1. Structurally, this is the complete binary tree with  $2^{\lceil \log_2 |U| \rceil} = 16$  leaves, cut off on the right side along the path between the root and the leaf corresponding to  $|U| - 1$ . It can be viewed as an uncompressed binary decision diagram [11], where each level of the tree is a decision based on one bit of an element, ordered from most to least significant. Each internal node  $t$  in the figure is labeled  $n_t/V_t$ , where  $V_t$  is the size of the subuniverse (the number of leaves) in the subtree rooted at  $t$ , and  $n_t$  is the size of the subset of  $S$  that falls in that subuniverse. Leaves, whose subuniverse size is always 1, are labeled only with subset size (0 or 1). We are concerned with representing only the part of the tree corresponding to nonempty subsets (solid-line edges in the figure), which for a sparse set is a relatively small part of the full tree. It can be viewed as a binary trie representing the elements of  $S$  (one can note, also the correspondence to the trie-oriented DST code [15]). In another interpretation, it resembles a wavelet tree [8], representing a string of unique symbols in increasing order.

The  $V_t$  values depend only on  $|U|$ . For the root, we have  $V_{root} = |U|$ . Let  $p$  be the root of a subtree of height  $h > 0$ . The subuniverse sizes of its left and right child are  $V_\ell = \min\{2^{h-1}, V_p\}$  and  $V_r = V_p - V_\ell$ .

We encode the tree by emitting  $|S| = n_{root}$  followed by, in a specific top-down order, subset size  $n_t$  for each node  $t$  that is a *left* child of some  $p$  with  $n_p > 0$ . The right sibling of  $t$  has subset size  $n_p - n_t$ , and hence does not need to be explicitly represented.

Traversing the tree top-down recursively narrows the ranges of subset sizes, similarly to the range-narrowing methods of section 2.2. We choose inorder (depth-first, left-to-right) traversal, although any well-defined top-down order would do. The values emitted in the figure example are those shown in frames, and the emit order is shown as numbers below the frames. As for their ranges, we clearly have  $0 \leq n_t \leq n_p$ ,

where  $p$  is the parent of  $t$ , but the range can often be bounded further. Let  $t$  be a node whose  $n_t$  is to be emitted and  $r$  its right sibling. Since  $n_r \leq V_r$  and  $n_t = n_p - n_r$ , we have  $n_t \geq n_p - V_r = n_p + V_t - V_p$ . Also, obviously  $n_t \leq V_t$ . Hence, the range of possible values of  $n_t$  is  $[\max\{0, n_p + V_t - V_p\}, \min\{n_p, V_t\}]$ .

Encoding the example in the figure begins with  $n_{root} = 6$  in the range  $[0, 11]$ . It then progresses with the left child of the root whose possible subset sizes, by the given computation, is between 3 and 6 (inclusive). Hence, the emitted value is  $5[3, 6]$ , and the rest of the sequence is  $2[1, 4]$ ,  $0[0, 2]$ ,  $1[1, 1]$ ,  $1[1, 2]$ ,  $0[0, 1]$ ,  $1[1, 1]$ ,  $1[1, 1]$ ,  $0[0, 1]$ , and  $1[1, 1]$ . For completeness, we include emitting in unit-size ranges (such as  $[1, 1]$ ) in the sequence, although it produces no bits in the encoding. Decoding works by tracing the same traversal as encoding, and can decode the emitted subset sizes thanks to the top-down order.

The simplest way to encode the  $n_t$  would be by a binary code, which makes the code somewhat similar to interpolative coding. However, a flat binary code corresponds to an implicit assumption of uniform distribution among the possible subset sizes, which would be a peculiar distribution to appear in practice. Hence, compression performance (row 6 in table 1) is not difficult to beat.

*Uniform Distribution* A more likely scenario would be uniform distribution among the *elements* in the subset. Let  $t$  be a non-root node whose subset size is to be encoded, and  $p$  its parent. Let  $s = V_t$  and  $f = V_p - V_t$ . The probability that  $n_t = m$  is that of  $m$  successes in  $n_p$  draws, *without* replacement, from a population of size  $V_p = s + f$ , whereof  $s$  individuals correspond to success and  $f$  to failure. This corresponds to *hypergeometric distribution* [9]. We have  $\Pr(n_t = m) = \binom{s}{m} \binom{f}{n_p - m} / \binom{s+f}{n_p}$ , and the expected value of  $n_t$  is  $n \times s / (s + f)$ . We can let this distribution decide probability ranges for arithmetic coding. Note that the distribution does not depend on the individual uniform probability of the elements.

Returning to the example in figure 1, let  $t$  be the root's left child, whose subset size is 5. We have  $\Pr(n_t = 5) = \binom{8}{5} \binom{3}{6-5} / \binom{8+3}{6} \approx 0.36$ . Summing over the possible range of set sizes in this case yields  $\sum_{m \in [3, 6]} \Pr(n_t = m) = 1$ , as we would expect.

In computing probability ranges for  $n_{root} = |S|$ , the hypergeometric distribution is not useful, since there are no known  $s$  and  $f$ . Instead, we could ideally assume that the probability for  $S$  having  $m$  elements is that of  $m$  successes in  $|U|$  draws, where the success probability is  $p = \Pr(x \in S | x \in U)$ . This corresponds to the binomial distribution [9]:  $\Pr(|S| = m) = \binom{|U|}{m} p^m (1 - p)^{|U| - m}$ . It is to be expected that for a uniform distribution where the element probability  $p$  is known, encoding  $S$  using the binomial distribution for  $|S|$  and the hypergeometric distribution for each subsequent  $n_t$  yields the same total encoding length as the corresponding yes/no code, i.e.,  $-|S| \log_2 p - (|U| - |S|) \log_2 (1 - p)$  bits. In particular, setting  $p$  to  $1/2$  should encode any set in  $|U|$  bits.

In the general case where  $p$  is not known, we must resort to a cruder estimate. When the specific application contributes no additional information, assuming uniform distribution over set sizes is perhaps the most reasonable compromise, since it limits the penalty to  $\log_2 |U|$  bits per sets. Unless  $S$  is very small, this adds relatively little to the encoding size.

	<i>txt/8</i>	<i>txt/24</i>	<i>words</i>	<i>inverted</i>
1 <i>gap</i>	1.71	2.04	5.02	4.64
2 <i>gap w/o repl.</i>	1.53	1.96	5.02	4.55
3 <i>yes/no</i>	1.70	1.70	5.09	5.25
4 <i>interpolative</i>	1.65	2.16	5.43	4.82
5 <i>tournament</i>	2.03	2.55	5.37	5.05
6 <i>RSSS flat</i>	1.99	2.61	5.60	5.12
7 <i>RSSS hypergeom.</i>	1.53	1.96	5.02	4.55

**Table 1.** Encoding lengths (bits per element) for non-statistical codes.

Table 1 shows test results for the methods discussed up to this point. Encoding length is given excluding the representation of  $|S|$ , except for the yes/no code, which does not require that  $|S|$  is represented separately.

Note the similarity in performance patterns of RSSS with the range-narrowing codes, but where RSSS appears to have a consistent overhead. We defer a formal analysis to future work, but conjecture that the overhead is at least partly due to how RSSS often needs to explicitly encode zero-size subsets, which simply do not appear in the range-narrowing codes. Note also that none of the methods are clearly better than gap coding for this case. We have no indication of RSSS contributing to compression performance unless the uniform-probability assumption is surpassed by the use of statistics, which is what we turn to next.

### 3.3 Using Statistics

As an example of a statistical scenario, assume that we can store observations taken from a large set of typical sample sets  $D = \{S_1, \dots, S_{|D|}\}$ . As observed in section 3.1, maintaining counters for producing an individual estimate for each of the possible  $2^{|U|}$  sets is unrealistic. Instead, our approach is to maintain counters in the binary tree over  $U$  described in section 3.2. For each left child  $t$  in the tree, we maintain  $C_t = \sum_i n_{t,i}$ , where  $n_{t,i}$  denotes the value of  $n_t$  (defined in section 3.2) in processing sample set  $S_i$ . We obtain  $q_t = C_t/C_p$ , which, on encoding the  $n_t$  of a specific set, lets us estimate the expected value of  $n_t$  by  $q_t n_p$ .

Henceforth, we simply assume that correct  $q_t$  estimates are available, by some prior knowledge about the distribution of sets. For our performance measurements, we obtain  $q_t$  values by counting element appearances, across the set of test inputs for respective tests. This is not intended as a suggestion for practical use, but consider it a choice for testing the capability of our method to adopt probability ranges in accordance with known statistics.

We now consider the use of  $q_t$  to find a better estimate for  $\Pr(n_t = m)$  for each  $m \in [\max\{0, n_p - f\}, \min\{n_p, s\}]$  (where  $s$  and  $f$  defined as in section 3.2). It may seem reasonable to use the same hypergeometric distribution as for the uniform assumption, after replacing  $s$  and  $f$  with  $s' = \lfloor q_t(s + f) \rfloor$  and  $f' = (s + f) - s'$ . However, this would assign zero probability to some possible  $n_t$ . For instance, let  $t$  be the left child of the root

in figure 1, and say that statistics tells us that  $q_t = 0.35$ , which yields  $s' = \lceil 0.35 \times 11 \rceil = 4$  and  $f' = 11 - 4 = 7$ , and the probability range for the desired value  $n_t = 5$  is zero.

Instead, we consider the following options.

*Binomial Approximation and Case Exclusion* For large  $|U|$ , hypergeometric distribution (*without replacement*) approaches the corresponding distribution *with replacement*, i.e. binomial distribution. Using binomial distribution as an estimate makes it simple to incorporate  $q_t$ , by setting  $\Pr(n_t = m) = \binom{n_p}{m} q_t^m (1 - q_t)^{n_p - m}$ . The desired expected value of  $q_t n$  is retained, and all possible values are given nonzero probabilities.

However, this estimate assigns nonzero probability to all  $0 \leq m \leq n_p$ , which may include values smaller than  $n_p - f$  and larger than  $s$ . This is clearly wasteful. For example, for  $s = 5, f = 5, n_p = 7$ , more than 80% of the range is taken by the impossible cases  $m = 0, m = 1$ . We adopt the following strategy to adjust the values of  $m, n_p, s, f$  before encoding, to remove the correct number of unusable states:

1. If  $n_p > s$ , reassign, in order,  $d \leftarrow n_p - s$ ,  $n_p \leftarrow s$ , and  $f \leftarrow f - d$ .
2. Then, if  $n_p > f$ , reassign, in order,  $d \leftarrow n_p - f$ ,  $m \leftarrow m - d$ ,  $n_p \leftarrow f$ , and  $s \leftarrow s - d$ .

*Scaled Hypergeometric Approximation* One possibility for modifying the values of  $s$  and  $f$  to let  $s/(s + f) = q_t$  before applying  $\Pr(n_t = m) = \binom{s}{m} \binom{f}{n_p - m} / \binom{s + f}{n_p}$ , while maintaining nonzero and reasonable probabilities for all possible subset sizes, is to scale up linearly. This maintains at least some of the *without replacement* property of the hypergeometric distribution, while the balance of left and right subrange is set to reflect statistical estimates. We have the following cases.

If  $q_t = 0$ , given that  $q_t$  values are to be trusted, we know with certainty that  $n_t = 0$ . If  $q_t$  merely reflects statistics over some training data, a probability range should be reserved for this case, the size of which may need adjusting dependent of the application. In our tests, we simply assume that the  $q_t$  values are correct (as they are in the experimental setting), and that  $n_t$  does not have to be explicitly represented.

If  $q_t = 1$ , we have, analogously, that  $n_t = \min\{n_p, s\}$ .

If  $s/f \geq q_t/(1 - q_t)$ , reassign  $f \leftarrow \lceil s(1 - q_t)/q_t \rceil$ .

If  $s/f < q_t/(1 - q_t)$ , reassign  $s \leftarrow \lceil f q_t / (1 - q_t) \rceil$ .

The transform may appear as somewhat ad hoc, but experiments indicate good performance, in particular when prepended with the *case exclusion* transform described above.

*Non-Central Hypergeometric Distribution* A less ad-hoc way of adjusting the hypergeometric distribution for the statistical case is to employ Wallenius' *non-central hypergeometric distribution* [5], henceforth referred to as NCHG. NCHG is a generalization of the hypergeometric distribution where a weight  $w$  introduces a bias between success and failure states. We set  $w = f/s \times q_t/(1 - q_t)$  for a suitable bias.

Computationally, NCHG is considerably more challenging than the previously listed distributions. We are aware of no closed form to produce the desired probability ranges exactly. For the *txt* data sets, we compute values of 1.14 and 1.62 bits per element, the most successful for the small sets, but we have been unable to compute the probabilities for the moderate-sized sets. Most likely, this is too inefficient to be considered.



		<i>words</i>	<i>inverted</i>
1	RSSS <i>binomial</i>	3.54	3.26
2	RSSS <i>rescaled hg</i>	3.48	3.22

**Table 2.** Encoding lengths (bits per element) of codes using element probabilities.

		<i>A</i>	<i>B</i>	<i>C</i>
1	<i>interpolative</i>	8.63	5.96	1.43
2	<i>tournament</i>	7.08	12.17	0.91
3	RSSS <i>binomial</i>	1.40	1.39	6.96
4	RSSS <i>rescaled hg</i>	1.39	1.37	8.26

**Table 3.** Encoding lengths (bits per element) for generated extreme sets.

Table 2 shows measurements for the binomial and rescaled hypergeometric distributions. As expected, the improvement from table 1 is significant.

### 3.4 Extreme Element Distributions and Universe Permutation

We now study the behavior on some elaborate variations in input data. First, consider the extreme case where  $S$  consists only of numbers divisible by  $k$ , and this is predicted correctly by the values of  $q_i$ . With  $k = 100$  and  $|U| = 10000$ , we get the results in column A of table 3.

This extreme can partially explain how RSSS captures properties that range-narrowing codes do not: the encoding length is zero for the lower  $\sim \log_2 k$  levels. To the range-narrowing codes elements are spread out evenly on each recursion depth. The behavior is somewhat similar even if the  $|U|/k$  elements that are the only ones to appear are randomly distributed across  $U$  (column B). (Tournament coding degenerates when the maximum element is near the low end of the range.) On the other hand, if the  $|U|/k$  elements lie only on one end of  $U$ 's range, we have the most skewed distribution, and the most compressible case for all the recursive methods, shown in column 3.

Finally, with access to global element probabilities, a simple modification of encoding is to permute the enumeration of  $U$  in probability order. As seen by comparing tables 4 and 1, this has a strong effect on range-narrowing codes. Also RSSS benefits somewhat compared to table 2, since skewness in the subset size distribution increases on higher levels of the tree.

## 4 Conclusion and Future Research

We have demonstrated several ways of exploiting statistical knowledge of elements when compressing integer sets, which has opened a number of paths for future research. The methods for which we have the strongest indication of good compression performance are still rather crude in some aspects. For instance, permutation based on

	<i>words</i>	<i>inverted</i>	
1	<i>interpolative</i>	2.68	2.78
2	<i>tournament</i>	2.79	2.95
3	<i>RSSS binomial</i>	2.92	2.72
4	<i>RSSS rescaled hg</i>	2.97	2.81

**Table 4.** Encoding lengths (bits per element) with alphabet permuted in probability order.

global probability, explored in the final section above, still does not address the issue of *context*, element probabilities conditioned on the presence of other elements. Consider, for example, the case where the global element probabilities are all approximately the same, but where there is a strong correlation leading certain elements scattered across the range of elements to frequently appear in the same sets. Permuting the range in probability order clearly does not capture such a regularity, and although the most simple cases would be easy to handle, using a general method to detect correlations, e.g., min-wise hashing [2], for probabilistic modeling of set compression, is nontrivial, and an interesting topic for future research.

Other areas to explore include set compression on a wider application area than integer strings. Recursive subset-size encoding may be applicable also for bitstrings of non-homogeneous or unlimited lengths.

## References

1. Boldi, P., Vigna, S.: Codes for the world wide web. *Internet Mathematics* 2(4), 407–429 (2005)
2. Broder, A.: On the resemblance and containment of documents. In: *Compression and Complexity of Sequences 1997. Proceedings.* pp. 21–29 (1997)
3. Cover, T.: Enumerative source encoding. *Information Theory, IEEE Transactions on* 19(1), 73–77 (1973)
4. Elias, P.: Universal codeword sets and representations of the integers. *IEEE Transactions on Information Theory* 21(2), 194–203 (Mar 1975)
5. Fog, A.: Calculation methods for wallenius’ noncentral hypergeometric distribution. *Communications in Statistics – Simulation and Computation* 37(2), 258–273 (2008)
6. Golomb, S.W.: Run-length encodings. *IEEE Transactions on Information Theory* 12(3), 399–401 (Jul 1966)
7. Gripon, V., Rabbat, M., Skachek, V., Gross, W.J.: Compressing multisets using tries. In: *Proc. IEEE Information Theory Workshop.* pp. 647–651 (Sep 2012)
8. Grossi, R., Gupta, A., Vitter, J.S.: High-order entropy-compressed text indexes. In: *Proc. ninth Ann. ACM–SIAM Symp. Discrete Algorithms.* pp. 841–850 (2003), <http://dl.acm.org/citation.cfm?id=644108.644250>
9. Jaynes, E.: *Probability Theory – The Logic of Science.* Cambridge University Press (2003)
10. Larsson, N.J.: Considerations and algorithms for compression of sets. In: *Proc. IEEE Data Compression Conf.* p. 503 (2013), poster abstract
11. Lee, C.: Representation of switching circuits by binary-decision programs. *Bell System Technical Journal* 38, 985–999 (1959)
12. Moffat, A.: Compressing integer sequences and sets. In: Kao, M.Y. (ed.) *Encyclopedia of Algorithms*, pp. 178–183. Springer-Verlag (2008)

13. Moffat, A., Stuiver, L.: Binary interpolative coding for effective index compression. *Information Retrieval* 3, 25–47 (2000)
14. Navarro, G., Mäkinen, V.: Compressed full-text indexes. *ACM Comput. Surv.* 39(1) (Apr 2007), <http://doi.acm.org/10.1145/1216370.1216372>
15. Reznik, Y.A.: Coding of sets of words. In: *Proc. IEEE Data Compression Conf.* pp. 43–52 (Mar 2011)
16. Rubin, F.: Arithmetic stream coding using fixed precision registers. *IEEE Trans. Inf. Theory* IT25(6), 672–675 (Nov 1979)
17. Shannon, C.E., Weaver, W.: *The Mathematical Theory of Communication*. The University of Illinois Press (1949)
18. Siebes, A., Vreeken, J., van Leeuwen, M.: Item sets that compress. In: *Proc. SIAM Conference on Data Mining*. pp. 393–404 (2006)
19. Teuhola, J.: Tournament coding of integer sequences. *The Computer Journal* 52(3), 368–377 (2009), <http://comjnl.oxfordjournals.org/content/52/3/368.abstract>
20. Teuhola, J.: Interpolative coding of integer sequences supporting log-time random access. *Information Processing & Management* 47(5), 742–761 (2011)
21. Witten, I.H., Moffat, A., Bell, T.C.: *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, second edn. (1999)
22. Zobel, J., Moffat, A.: Inverted files for text search engines. *ACM Comput. Surv.* 38(2) (Jul 2006)